DOCUMENT RESUME

ED 333 047 TM 016 616

AUTHOR Linacre, John Michael
TITLE Constructing Measurement with a Many-Facet Rasch Model.
PUB DATE Apr 91
NOTE 16p.; Paper presented at the Annual Meeting of the American Educational Research Association (Chicago, IL, April 3-7, 1991).
PUB TYPE Reports - Evaluative/Feasibility (142) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Equations (Mathematics); Evaluators; Item Response Theory; *Mathematical Models; Psychometrics; Rating Scales; *Test Construction; Test Reliability; Test Validity
IDENTIFIERS *Rasch Model

ABSTRACT

The psychometric objections to using essays and other subjective tests for measurement can be overcome by a many-facet Rasch model. This model enables judge-awarded grades to be transformed from their arbitrary, local, non-linear rating scale form into linear measures with explicit generalizable meaning of specifiable reliability (standard error) and validity (fit). Flexible judging plans and accommodation for missing data simplify application. Quality control over the judging processing is readily available through meaningful fit statistics. The following is a sample many-facet Rasch model for an essay examination, using a rating scale of M + 1 ordered response categories with "0" labeling the lowest and "M" labeling the highest: B(sub n) is the ability of examinee "n"; D(sub i) is the difficulty of item "i"; C(sub j) is the severity of judge "j"; and F(sub k) is the difficulty of the step up from category k-1 to category k, and k=1,M. A maximum likelihood estimate is obtained when the expected marginal score for observations in which the parameter participates is set equal to the observed score. It is concluded that the many-facet model can be applied to any testing situation in which the constituent components can be conceptualized to combine additively to govern the raw rating category responses. Six figures are included. (Author/SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☑ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
JOHN M. LINACRE
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

ED333047

Constructing measurement with a many-facet Rasch model

by

John Michael Linacre

MESA Psychometric Laboratory
Department of Education
University of Chicago

Paper presented at
American Educational Research Association Annual Meeting
Chicago, Illinois
April 1991

TM016616

BEST COPY AVAILABLE

Abstract
The psychometric objections to using essays and other subjective tests for measurement can be overcome by a many-facet Rasch model. This model enables judge-awarded grades to be transformed from their arbitrary, local, non-linear rating scale form into linear measures with explicit generalizable meaning of specifiable reliability (standard error) and validity (fit). Flexible judging plans and accommodation for missing data simplify application. Quality control over the judging processing is readily available through meaningful fit statistics. The many-facet model can be applied to any testing situation in which the constituent components can be conceptualized to combine additively to govern the raw rating category responses.

Key words: Rasch Measurement, FACETS model, Rating Scales, Judge-awarded grades, essay grading, subjective tests.

"Subjective" versus "objective" tests:

The rush to "objective", multiple-choice question (MCQ), tests in the 1920's was driven by dissatisfaction with "subjective", judged-rated, tests. Objective tests were intended to control intrusions of undesirable variance into subjective test scores. But, in the 1980's, the testing community began to realize that what was needed is not "objective testing" but rather "objective measurement". The reevaluation of subjective tests in the light of objective measurement opens a new field of possibilities.

Ruch (1929) summarized the drawbacks to "subjective tests":

1. Subjectivity of scoring lowers reliability.
2. Sampling must be limited to a small number of broad questions.
3. Time required to write lengthy answers is excessive.
4. These examinations encourage bluffing.

His first objection is our primary concern here. The importance of his last three objections depend on the intention, construction and application of the subjective test. Indeed, the success of test-taking strategies to bluff on MCQ tests is one of the motivations behind the reintroduction of subjective tests.

Conventional attempts to model judging

Studies of scoring subjectivity have found that "there is as much variation among judges as to the value of each paper as there is variation among papers in the estimation of each judge" (Ruggles 1911). But any difference among judges is a threat to fairness because raw score depends on which judge rates an examinee. This is also true of variance introduced by any other factor, such as the time of day the ratings are awarded.

Since differences in judge severity can account for as much variance in ratings as differences in examinee ability (Cason and Cason 1984), an obvious and widely attempted correction

for judge behavior is to deduct the mean value of all ratings given by a judge from his individual ratings in hope of obtaining a "judge-free" rating. This fails because:

1) All judges are required to rate all examinees on all items, a design which is impractical in any large-scale testing situation (see Figure 3). Substituting partial sampling designs (Braun 1988) lessens the judging load, but introduces daunting administrative requirements.

2) The stochastic aspect of the judging process remains unrecognized and unmanaged. Adjustments by averaging and subtracting do not control the effects of judge variation.

3) The non-linearity of the initial rating scale is overlooked. Ratings originate on an ordinal, not an interval, scale. As Figure 1 illustrates,
   i) the highest and lowest categories represent infinite ranges of performance above and below the intermediate categories.
   ii) the ranges of performance represented by intermediate categories depend on how their labels are interpreted by judges. They are never equal.

4) Judge idiosyncracies are undiagnosed and uncontrolled. This means that the validity of the examination is unsupervised.

5) Measures for examinees cannot be produced which are statistically independent of the local details of the examination and hence generalizable beyond the examination.

Though attempts have been made to overcome these difficulties with conventional approaches, none has succeeded (Linacre 1989).

The many-facet Rasch model

These concerns can be overcome with a many-facet Rasch model. The axioms underlying the two-facet Rasch model (Rasch 1960/1980) can be extended to tests of many facets (Linacre 1989):

a) the impact of each element of each facet on the test situation is dominated by a single parameter with a value independent of all other parameters within the frame of reference. (Single parameterization is necessary if examinees are to be arranged in one order of merit, or items indexed by difficulty on an item bank).

b) these parameters combine additively - they share one linear scale.

c) the estimate of any parameter is dependent on the accumulation of all ratings in which it participates but is independent of the particular values of any of those ratings.

These axioms are the necessary and sufficient requirements for constructing a linear measurement system from observed data. The degree to which this construction is useful and valid is measured by statistics quantifying the fit of the data to the measurement model (Wright & Masters 1982).

A many-facet Rasch model for an essay examination, using a rating scale of M+1 ordered response categories with "0" labeling the lowest and "M" labeling the highest, is

(1)

where

B_n is the ability of examinee n
D_i is the difficulty of item i
C_j is the severity of judge j
F_k is the difficulty of the step up from category k-1 to category k, and k=1,M.

A maximum likelihood estimate for each parameter is obtained when the expected marginal score for observations in which the parameter participates is set equal to the observed score. Missing observations can be ignored in this estimation as is done in the computer program FACETS (Linacre 1988).

The non-linear rating scale

Equation (1) represents the relationship between the ordered categories of a rating scale and the latent performance continuum by means of an ogive which satisfies both the theoretical requirements for measurement and the functional form of the rating scale defined by the judges through their use of it. The unequal widths of the performance ranges corresponding to the intermediate categories are parameterized and estimated by the F_k terms. The infinite performance ranges at the extremes of the scale are mapped into the corresponding finite top and bottom categories. Figure 2 illustrates such an ogive.

The stochastic nature of judging

The many-facet Rasch model also handles the second reason conventional methods fail. The probability of a rating in any category is modelled explicitly, as is evident from equation (1). This leads to equation (2) for an expected score:

(2)

where

X_{nij} = the observed rating

E_{nij} = Σ k P_{nijk}, the expected score, with sum over k=0,M

and the modelled "error" variance is

$$(S_{nij})^2 = \sum_{k=0}^{M} (k - E_{nij})^2 \; P_{nijk} \qquad (3)$$

Let us compare this with the conventional model for ratings, in which rating category labels are treated as though they were linear measures (c.f. Braun 1988):

$$X_{nij} = E_{nij} \pm S'_j = B'_n - D'_i - C'_j \pm S'_j \qquad (4)$$

where

E'_{nij} is the expected score based on a linear regression on the elements of the facets

$(S'_j)^2$ is the unexplained error variance left by the regression. The size of S'_j is determined empirically and usually viewed as judge-dependent.

Even beyond the awkward modelling of ordinal, non-linear, rating labels as though they were linear measures, there are important differences between S_{nij} and S'_j.

In the Rasch equation, $(S_{nij})^2$ models the expected error variance for a particular observation. The estimation procedure constructs a linear scale on which the parameter estimates meet the condition that for each parameter, the observed and expected marginal scores coincide. Once this condition is met, the modelled (i.e. expected) values of S_{nij} are also known. The two right-most terms of (2) are thus obtained explicitly from the model equation (1) and the parameter estimates.

This enables a detailed examination of the data for fit to the model. Model specification of the magnitude of expected error provides an expliciit and reproducible frame of reference against which to evaluate each batch of data. Not only too much, but also too little, observed "error" variance threatens the validity of the measurement process, and motivates investigation, diagnosis and remediation of the specific measurement problems discovered to relate to specific observations. The relationships between the modelled "error" variances and the observed "error" variances (sums of squared residuals) can also be used as partial and global tests of fit of data to model (by an extension of Windmeijer, 1990).

$(S'_j)^2$, in contrast, is an unmodelled residual comprising a judge's unexplained and undesired "error" variance. The optimum value is zero, but this can never be observed in a non-trivial situation. Any amount greater than zero threatens validity. Thus, in a similar context, "the widespread use of such items in standardized tests depends on whether some degree of scoring error, however small, can be accepted" (Bennett et al. 1990). This error variance is often compared to the variance of a judge's observed ratings, leading to an uncontrolled comparison between the randomness of a judge's ratings and the spread of the abilities of the examinees who happen to have been rated.

The judging plan

The only requirement on the judging plan is that there be enough linkage between all elements of all facets that all parameters can be estimated within one frame of reference without indeterminacy. An example of lack of linkage and consequent indeterminacy is a plan in which judge panel A grades only boys and judge panel B grades only girls, because then a relatively good performance by one gender can be attributed either to higher ability or to more lenient judges.

Complete data, as shown in Figure 3, meets the linkage requirement and enables more precise measurement, but is not required. All that is required is a network of overlap of examinees, judges and items. A simple way this can be obtained is by having groups of judges rate some examinees on all items. This type of plan is shown in Figure 4.

Judging, however, is time-consuming and expensive. It is more desirable to minimize the work by arranging for each performance to be judged only once. Even under these circumstances, the statistical requirement for overlap can usually be met rather easily. If each examinee writes several essays and all essays are shuffled together randomly, overlap can be obtained by having each judge grade whichever essay happens to come next on the pile. Each judge grades as many essays as time and speed allow. But each essay is graded only once. Nevertheless, by the end of the judging session, many examinees will have been rated by more than one judge, but on different essays, and many essay topics will have been rated by more than one judge, but for different examinees. An example of this type of plan, but under slightly stricter rules, is shown in Figure 5.

Generalizability of results

The category labels of a rating scale are not only arbitrary and non-linear, but also local to the design of the particular examination. The implications of this may be masked when all examinees are rated on the same items by the same judges in one testing session, but they are immediately apparent when examinees face different testing situations. Quantitative comparison requires a frame of reference in which it no longer matters which examinee is rated by which judge on which item in what session. The many-facet Rasch model enables such a framework to be constructed (Stahl 1991) (see Figure 6). In contrast, generalizing

even the familiar SAT and ACT tests using conventional approaches has proved unmanageable (Wainer et al. 1985).

Control of judge idiosyncracy

Judge training is required to develop a shared understanding of a rating scale and a uniform perspective on the challenge applied by the test items. It is claimed that "Subjectivity of marking may be reduced about one-half by the adoption of and adherence to a set of scoring rules when essay examinations are to be graded" (Ruch 1929). Conventionally, training has been further aimed at obtaining unanimity across judges about the rating to be awarded to particular performances on particular items. This "idealistic" attempt to produce identical, and hence exchangeable, judges has met with little success. "Judges employ unique perceptions which are not easily altered by training" (Lunz et al. 1990). No entirely successful large-scale judge training program has been reported. There are many situations in which judge training is given little or no attention, (e.g. supervisor rating of subordinates), or is been discovered to have been ineffective.

It is always essential to monitor the quality of the ratings being awarded and to direct each judge's attention to those areas in which there is doubt. An advantage of the facets measurement model is that judge self-consistency, rather than unanimity, is now the aim. On this basis, unexpectedly harsh or lenient ratings can be identified, and also each judge's "biases" relating to any particular items, groups of examinees, or the like, can be quickly revealed. This has two benefits. First, unacceptably idiosyncratic ratings can be treated as "missing" without disturbing the validity of the remainder of the analysis. Second, precise feedback to each judge about specific questionable ratings or rating patterns can foster improvements in the judging process.

Further measurement models

The many-facet measurement model can be expressed in many forms to meet the requirements of specific testing situations. Other forms include:

An item-scale model, in which each item is constructed with its own rating scale:

$$\log\left(\frac{P_{nijk}}{P_{nijk-1}}\right) = B_n - D_i - C_j - F_{ik} \qquad (5)$$

where

B_n is the ability of examinee n
D_i is the difficulty of item i
C_j is the severity of judge j
F_{ik} is the difficulty of the step from category k-1 to category k of the scale unique to **item i**, and $k = 1, M_i$

A judge-scale model, in which each judge uses his own interpretation of the rating scale:

$$\log\left(\frac{P_{nij k}}{P_{nij k-1}}\right) = B_n - D_i - C_j - F_{jk} \qquad (6)$$

where

B_n is the ability of examinee n
D_i is the difficulty of item i
C_j is the severity of judge j
F_{jk} is the difficulty of the step from category k-1 to category k for **judge j**, and $k = 1, M_j$

A four-faceted model, in which each of the items is modelled to apply to each of a number of tasks:

$$\log\left(\frac{P_{nmijk}}{P_{nmijk-1}}\right) = B_n - A_m - D_i - C_j - F_k \qquad (7)$$

where

A_m is the difficulty of task m
B_n is the ability of examinee n
D_i is the difficulty of item i
C_j is the severity of judge j
F_k is the difficulty of the step from category k-1 to category k for judge j, and $k = 1, M$

Conclusion

The construction of a measurement system for essay grading and other "subjective" tests is practical and useful. Test constructors no longer need limit themselves to what little can be obtained from an "objective" test, but instead can devote their creative powers to designing tests which involve deeper and more relevant evidence of competence and yet which also produce objective measures.

References

Bennett R E, Ward W C, Rock D A, Lahart C (1990) Toward a framework for constructed-response items. Research Report RR-90-7. Princeton: Educational Testing Service.

Braun H I (1988) Understanding scoring reliability. Journal of Educational Statistics. 13(1) 1-18.

Cason G J, Cason C L (1984) A deterministic theory of clinical performance rating. Evaluation and the health professions 7:221-247

Linacre J M (1988) FACETS computer program for many-facet Rasch analysis. Chicago: MESA Press.

Linacre J M (1989) Many-faceted Rasch measurement. Chicago: MESA Press.

Lunz M E, Wright B D, Linacre J M (1990) Measuring the impact of judge severity on examination scores. Applied Measurement in Education 3(4), 331-345

Ruch G M (1929) The objective of new-type examination. Chicago: Scott, Foresman

Ruggles A M (1911) Grades and grading. New York: Teacher's College.

Stahl J (1991) Equating examinations that require judges. Paper presented at AERA Annual Meeting. Chicago.

Wainer H, Holland P W, Swinton S, Wang M H (1985) On State Education Statistics. Journal of Educational Statistics 10(4) 293-325.

Windmeijer F A G (1990) The asymptotic distribution of the sum of weighted squared residuals in binary choice models. Statistica Neerlandica 44(2) 69-78.

Wright B D, Masters G N (1982) Rating Scale Analysis. Chicago: MESA Press.

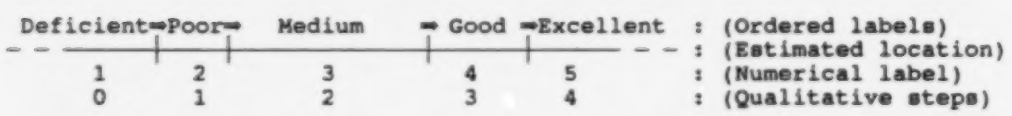
Rating categories are conceptualized as representing qualitatively distinct, but ordered, performance levels on an infinite continuum of performance.

The scale as it is printed:

```
Deficient    Poor     Medium    Good    Excellent : (Word label)
    1          2        3         4         5     : (Numerical label)
```

Equal spacing invites the rater to devote equal attention to each of the alternatives. Conventional analysis mistakes the arbitrary numerical labels for measures and so ignores the non-linear functioning of the scale.

The scale as it actually functions:



The range of performance corresponding to each of the ordered categories is discovered empirically by how the raters behave. The extreme ranges are always infinite. It is the functioning scale which defines the measures.

Figure 1. Conventional and measurement perspectives on rating scales.

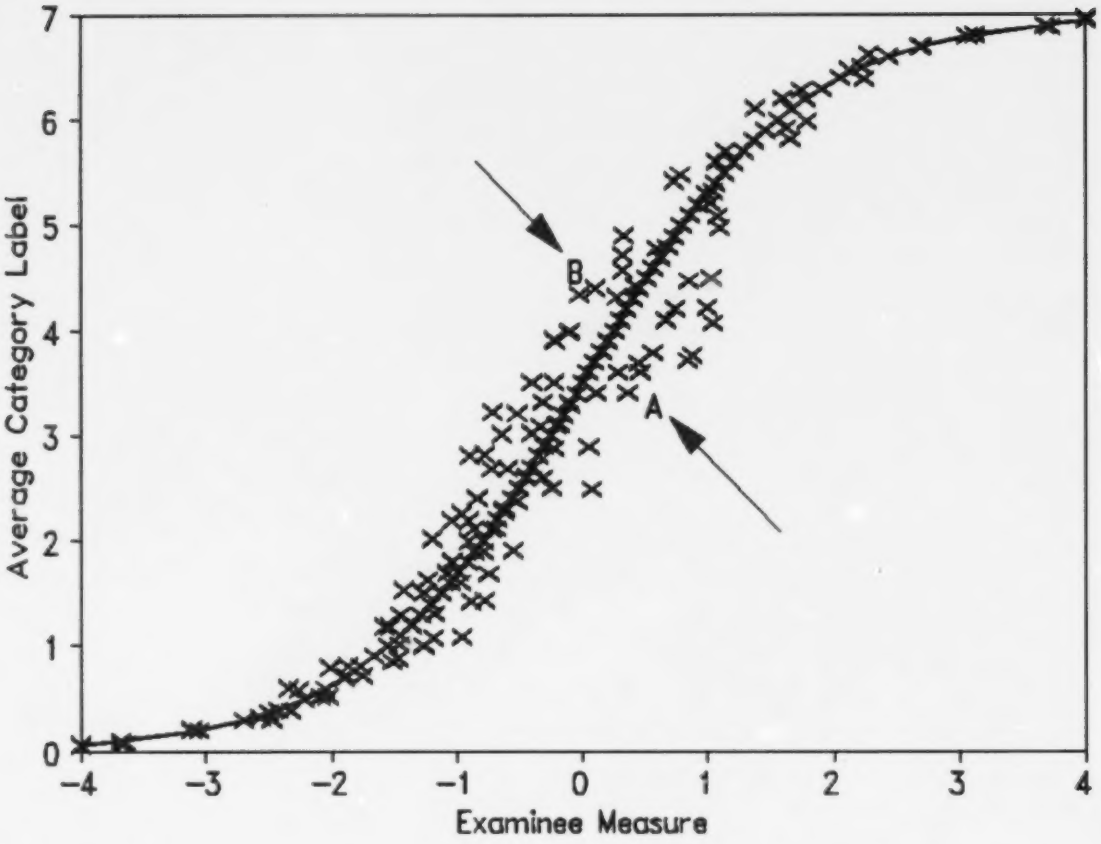


Figure 2. Average category labels for examinee performance plotted against estimated logit measures.

The solid ogive traces the raw score to measure conversion when all judges rate all examinees on all essays. Each X represents a conversion when only some of the judges rate an examinee's performance. Examinee A has a higher measure, but lower average rating label than Examinee B, because A happened to be rated by more severe judges than B.

Judge	1	2	3	4	5	6	7	8	9	10	11	12
Essay	ABC	ABC	ABC	ABC	ABC	ABC	ABC	ABC	ABC	ABC	ABC	ABC
Person												
1	553	686	877	687	777	685	565	667	586	567	776	696
2	454	542	445	534	334	344	433	526	444	445	533	534
3	434	544	343	555	433	544	563	443	554	454	443	343
4	345	426	232	545	445	225	464	456	642	446	445	335
5	443	548	656	545	657	448	558	466	464	448	547	348
6	544	846	843	565	633	367	788	673	666	566	564	454
7	545	665	454	667	755	646	773	785	874	565	745	447
8	553	763	655	675	775	653	773	656	784	576	573	574
9	343	643	643	645	534	523	665	674	753	546	545	765
10	564	766	884	776	655	667	875	778	778	667	649	888
11	535	524	537	544	545	435	546	557	326	446	456	334
12	436	644	444	546	666	555	574	445	745	356	763	676
13	445	486	657	566	246	366	368	448	467	348	569	349
14	446	533	333	344	545	343	463	353	354	346	462	363
15	548	855	743	746	766	656	665	765	854	666	862	844
16	644	653	547	545	643	454	556	467	666	447	558	667
17	414	817	625	628	536	518	425	618	717	627	639	436
18	334	655	443	445	243	473	445	747	654	445	435	334
19	747	745	837	756	755	847	664	688	737	656	847	938
20	443	666	735	556	557	557	588	667	666	557	476	488
21	242	443	336	465	245	243	263	245	441	253	342	254
22	564	765	747	666	864	577	667	576	667	557	667	785
23	446	566	753	646	444	565	475	388	576	557	557	776
24	332	422	334	433	322	214	423	223	323	313	233	223
25	543	664	544	657	646	544	454	448	547	545	456	464
26	644	764	955	756	545	658	655	867	776	646	756	885
27	342	346	334	344	346	234	256	256	345	345	256	253
28	343	463	335	334	465	573	341	475	442	243	462	272
29	433	444	323	446	334	333	235	336	423	336	323	343
30	542	564	244	655	445	224	546	575	645	446	432	555
31	325	514	313	425	315	314	334	225	525	314	324	314
32	644	744	445	545	533	553	567	584	664	447	556	364

Figure 3. **Complete judging plan.**
The ideal judging plan for conventional analysis. Under Rasch analysis, it provides precise measures of all parameters in the shared frame of reference. (Information provided by H. I. Braun contributed to this figure).

Judge	1	2	3	4	5	6	7	8	9	10	11	12
Essay	ABC	ABC	ABC	ABC	ABC	ABC	ABC	ABC	ABC	ABC	ABC	ABC
Person												
1	553	686										
2		542	445									
3			343	555								
4				545	445							
5					657	448						
6						367	788					
7							773	785				
8								656	784			
9									753	546		
10										667	649	
11											456	334
12	436											676
13	445						368					
14		533						353				
15			743						854			
16				545						447		
17					536						639	
18						473						334
19	747			756								
20		666			557							
21			336			243						
22				666			667					
23					444			388				
24						214			323			
25							454			545		
26								867			756	
27									345			253
28	343									243		
29		444									323	
30			244									555
31	grading performed by any available judges											
32	grading performed by any available judges											

Figure 4. **"Rotating test-book" judging plan.**
Parameters are linked into one frame of reference through ratings which share pairs of parameters (common persons, common examinees or common judges). Accidental omissions or unintended ratings amend the judging plan, but do not threaten the analysis. Measures are less precise than with complete data because fewer observations are made.
(Information provided by H. I. Braun contributed to this figure).

Judge	1	2	3	4	5	6	7	8	9	10	11	12
Essay	ABC	ABC	ABC	ABC	ABC	ABC	ABC	ABC	ABC	ABC	ABC	ABC
Person												
1							6	7	5			
2				5	3					5		
3	4									5	3	
4		2				5				4		
5					7					4	5	
6	5	6									6	
7					5		3					4
8		6						6		5		
9					3		6	4				
10				7	5				7			
11									6		4	3
12	4			6								7
13	4			6					4			
14	6	3					4					
15	4								4	6		
16		6	4								8	
17			2	6	6							
18						4	4					4
19		7		6		4						
20			7	5					6			
21						2	6			3		
22			7		8				6			
23		6	7					8				
24						2	3					2
25						4	4				4	
26									7	6		8
27			3					2			6	
28	4	3										2
29			3					3				3
30	2							5			3	
31				2		4		2				
32			5		5	5						

Figure 5. **Minimal-effort judging plan.**
Each of the 32 examinees' three essays is rated by only one judge. Each of the 12 judges rates 8 essays, including 2 or 3 of each essay type. The examinee-judge-essay overlap enables all parameters to be estimated unambiguously in one frame of reference.
Assignment of essays to judges was by a simulated "random pile" of essays with the constraints that each essay be rated only once; each judge rate an examinee once at most; and each rater avoid rating any one type of essay too frequently. The cost of minimal data collection is low measurement precision. (Information provided by H. I. Braun contributed to this figure).

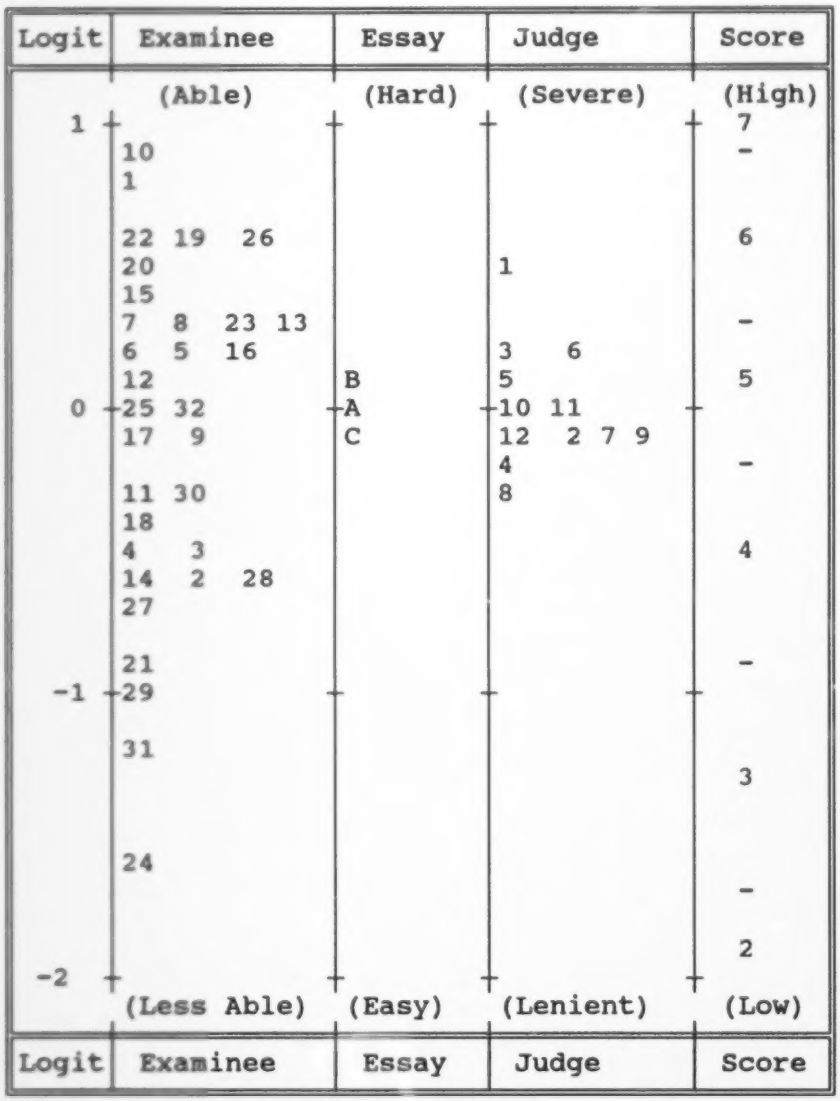


Figure 6. Results of a many-facet Rasch analysis. Examinee, essays and judges are measured on one common linear frame of reference. The expected scores (in rating points) are shown for examinees facing essays of 0 logit difficulty and judges of 0 logit severity. Other expected scores are obtained by indexing the score scale at (examinee - essay - judge) logits.
(Information provided by H. I. Braun contributed to this figure).